Object tracking using level set and MPEG 7 color features

Oussalah, M.; Shabash, M.

DOI:
10.1109/IPTA.2012.6469575

Document Version
Early version, also known as pre-print

Citation for published version (Harvard):
Oussalah, M & Shabash, M 2012, Object tracking using level set and MPEG 7 color features. in *Image Processing Theory, Tools and Applications (IPTA), 2012 3rd International Conference.* pp. 105-110, 2012 3rd International Conference on Image Processing Theory, Tools and Applications (IPTA), United Kingdom, 15/10/12. https://doi.org/10.1109/IPTA.2012.6469575

Object Tracking Using Level Set and MPEG 7 Color Features

M. Oussalah and M. Shabash
University of Birmingham, School of Electronics, Electrical and Computer Engineering
Edgbaston, Birmingham, B15 2TT, UK
M.Oussalah@bham.ac.uk

Abstract -- **This paper investigates the use of a binary level set-based approach for tracking vehicles in video sequences captured by camera systems. A combination of color and segmentation features including background subtraction and connected component labeling will be incorporated into the level set-based framework to achieve the tracking. Next, the retrieval task with respect to a given user's request image is achieved using a set of similarity indices constructed through MPEG-7 color features; namely, color histogram, dominant color descriptor and color layout descriptor. The performances of the system are evaluated in terms of precision and recall evaluations.**

Index Terms—**Level Set, background subtraction, tracking, image retrieval.**

I. INTRODUCTION

With the exponential increase of datasets captured by automated video systems, the task of tracking suspected objects becomes very expensive. This task is nowadays heavily related on human operators for manual analysis of the video sequences. Consequently, constructing an automated system for such purpose is highly recommended as it can be cost effective for many security-related regulators. However this task is challenged by the limitations in computer image analysis due to high computational cost and inherent restriction of image segmentation algorithms. This motivates the extensive research accomplished in recent years in this topic. The emergence of active contour like approaches in recent years [1-3] has contributed significantly to overcome many difficulties pervading standard segmentation approaches. Active contour approaches can be classified into the parametric models, and the geometric models that are based on level set methods. The snake model [4] and the related methods, e.g., [5,6], use the minimization of an energy functional which is a sum of an internal energy and an external energy. One shortcoming of the methods is that they cannot detect multiple objects. Geometric models based on level set functions can detect multiple objects due to their ability to adapt to the changes of topology. Some of the advancements in this approach are the geodesic contour models by Caselles, Kimmel, Sapiro [7], the geodesic region based method by Paragios-Dericbe [8] for texture image segmentation and visual motion tracking, and the Chan-Vese (C-V) model [9] which is based on the Mumford-Shah model [10]. The C-V model does not rely on the gradients of the intensity levels and thus is applicable for segmentation of certain images that do not have edges. Chan an d Vese also pointed out that the capture range could be extended by replacing the delta function with the magnitude of the gradient of the level set function.

Zhang et al. [11] proposed a new model based on the association of two models based edge and region: the geodesic and the Chan-Vese models. The objective of this combination is to exploit the advantage of the two types of models. The new obtained model is based on the statistical information inside and outside the contour to construct a region-based signed pressure force (SPF) function, which is able to control the direction of evolution of the curve. The authors also proposed a new selective binary and Gaussian filtering regularized level set (SBGFRSL) to implement the new proposed model. This method presents many advantages. First, it avoids the calculation of SDF. Second it avoids re-initialization. Third, it becomes more efficient than the traditional level set. Fourth, it has the property of selective local or global segmentation.

A typical scenario considered in this paper is a situation where a car needs to be spotted in a large area around a given city, an operator would usually have to go through each and every car that has entered the area, recorded by the surveillance cameras manually. But, in the case of automated surveillance systems, this task can easily be carried out by searching for the car according to the features of the particular car. The approach advocated in this paper makes use of binary level set based approach in light of [11] combined with appropriate background subtraction method to enhance the pre-processing stage. Next, a set of colour features suggested in MPEG 7 standards is employed to build a similarity index that allows us to compare detected images with the given request. The next section provides a rough description of the global approach.

978-1-4673-2584-4/12/$31.00 ©2012 IEEE

II. OVERALL APPROACH

As illustrated in Figure 1, the system consists of four main parts. At a begging, a snap of video was recorded and the frames of this video were extracted afterwards. In order to detect the vehicles, the frames were subjected to several procedures. Background subtraction is the first step of the detection procedures. It is used to aid the (binary) level set-based framework in recovering the shape of the vehicles. Connected component- labeling is employed in this work to detect joined regions. (i.e. non-target object-vehicles) in the frames. After detecting the vehicles successfully, different color features, employed in MPEG-7 color standard [12], were used for feature extraction. This includes colour histogram, dominant color and color layout. In order to recognize the vehicles, a matching method based on similarity evaluation is used on each color feature.

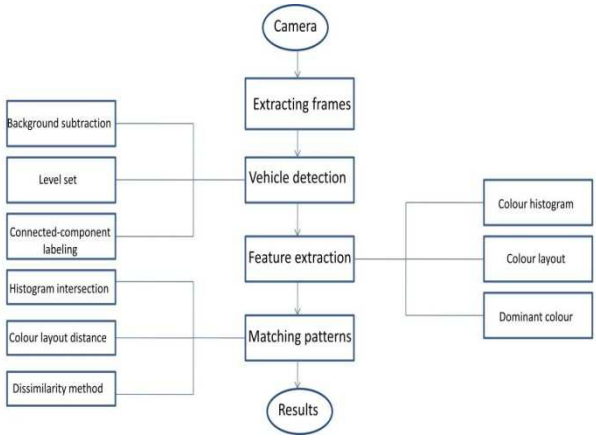

Figure 1. System overview

III. PRE-PROCESSING

A. Background Subtraction

Background subtraction is one of the most promising approaches for recognizing moving objects from a video sequence. Its main functionality is to separate the foreground objects from the background, in a sequence of video frames. It has been used in many computer-vision applications, such as traffic monitoring, gesture recognition for human-machine interfaces, and video surveillance. Pixel values that experience significant changes in their current frame, from the background, are considered to be moving objects. Among diverse method for this purpose, one distinguishes non-recursive techniques [13] that store a buffer of preceding video sequences, and then evaluate the background image that is anchored in the temporal variation of each pixel within the buffer. A simple example of this is the frame difference method. It is arguably considered to be the simplest technique of background subtraction. Mainly, the current frame is subtracted from the previous one, and a threshold of Ts is applied to the difference in pixel values. If the pixel has a greater value than Ts, it is considered to be a part of the foreground.

$$|frame_i - frame_{i-1}| > T_s$$

One advantage of this background modeling technique is that the background is highly adaptive. Since the background and the previous frame are related, the background modeling technique can adapt to changes in the background faster than any other method (at 1/fps to be precise). In addition, it is considered to be the most modest computational load. However, it can be noted that the major disadvantage of this approach is that the interior pixels can be considered a part of the background for objects with uniformly distributed intensity values. For this purpose, we adapted the approximated median filter put forward by McFarlane [14]. In this respect, if a pixel has a value that is larger than the corresponding background pixel, the condition is incremented by 1, and if the given pixel has a value that is less than the corresponding background pixel, that background will be decremented by one. This implies that the background eventually converges to an estimate where half of the input pixels are greater than the background and the other half are less than the background, in order to yield the approximate median. The frame rate, as well as the amount movement in the scene, determine the meeting time [13].

B. Connected Component labeling

Connected-component labeling is an application employed in detecting connected pixels in binary digital images, although color images and data with higher-dimensionality can also be processed. A group of pixels that are connected based on connectivity types are called an object. Connected-component analysis can process a variety of information when it is incorporated into a human-computer interaction interface or image recognition system.

The number of objects that can be found in an image might be affected by the type of neighborhood-checks. If the algorithm checks the labels of pixels that are on the West, West-North, North, North-East, East, East-South, South, and South-West of a current pixel, then this type is called an 8-neighborhood-check. If the algorithm checks the labels of pixels that are on the edges, i.e. west of a current pixel, then this type is called a 4-neighborhood-check, see for instance [15,17].

C. Feature extraction

A subset of the color descriptors has been standardized by MPEG-7 [12] that provides a content representation, and the standard for fast and effective searching of objects of interest.

1) Color space descriptor

Color space descriptors were utilized by a color histogram that represented a HSV color space quantized to 64 bins. The RGB image of the vehicle was converted to an HSV color space. Then, the histogram of the image was taken in 64 bins.

2) Dominant color descriptor

Compared to the normal histogram based descriptors, the representative colors are evaluated from each image or a region of that image rather than being fixed in the color space. This makes the representation of the color compact and accurate

Generally, the dominant color descriptor includes the number of dominant colors (N). Each dominant color descriptor (N) is defined by the following equation

$F= \{\{ c_i , p_i \}\}, (i=1 ,.....N)$

Where c_i represents the color components (i.e. a 3-D color component in the RGB color space) and p_i represents the percentage of pixels or the portion of pixels corresponding to the color component c_i.

In order to extract the dominant colors from an image, a color quantization has to be evaluated. This involves converting images from RGB to the HSV color space, and then quantized at 166 colors.

3) Color layout descriptor

The color layout descriptor (CLD) is achieved by implementing the discrete cosine transform (DCT) on a 2-D array of the representative colors in the color space Y/Cb/Cr. The approach to extract the CLD in an image can be explained in the following four phases [19]:

- Dividing the input frame to 64 blocks to insure the scale invariance.
- Evaluating the representative color for each block. Average of the pixels colors method is suggested for this task.
- The result in 8x8 icon image is converted from the RGB to Y/Cb/Cr color space, then an 8x8 DCT is implemented to the icon image so three set of 64 DCT-coefficients are attained.
- Last stage is done by scanning these coefficients using Zig-zag scanning technique.

IV. BINARY LEVEL SET

The level set method is one of the most useful tracking methods. The sole purpose of it is to recover the shape of the interested object and isolate it from its background. This approach provides several advantages over the parametric active counters. First, the level set function remains on a fixed grid that allows efficient numerical schemes. Second, it may merge or break during the evolution; as a result, the topological changes will be handled [3, 18]. In our approach we adopt the model proposed in [4], which is based on the association of the geodesic active contour and the Chan-Vese model. This allows us to exploit the advantages of local segmentation in order to choose an object in the image and track it without the need to segment the entire image. The edge stopping function (ESF) in the geodesic model is replaced by the SPF. Let Ω be a bounded open subset of $\Re^2$ and I is the given image.

Let $C(q):[0,1]\to\Re^+$ be a parameterized planar curve in Ω. With the level set method, we assume:

$$\begin{aligned} &C=\{x\in\Omega:\varphi(x)=0\}\\ &inside\ C=\{x\in\Omega:\varphi(x)<0\}\\ &outside\ C=\{x\in\Omega:\varphi(x)>0\}\end{aligned} \tag{1}$$

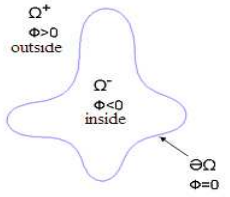

Figure 2. initial curve and 0-level curve

The level set formulation of the geodesic active contour presented in [11] is given as follow

$$\frac{\partial\phi}{\partial t}=g|\nabla\phi|\left(div\left(\frac{\nabla\phi}{|\nabla\phi|}\right)+\alpha\right)+\nabla g\,.\,\nabla\phi \tag{2}$$

where g is the edge stopping function (ESF), α is the balloon force which controls the contour shrinking or expanding and ∇ is the gradient operator.

Also the level set formulation of the Chan-Vese active contour is given by expression (3).

$$\frac{\partial\phi}{\partial t}=\delta(\phi)\left[\mu\nabla\left(\frac{\nabla\phi}{|\nabla\phi|}\right)-\upsilon-\lambda_1(I-c_1)^2+\lambda_2(I-c_2)^2\right] \tag{3}$$

where $\mu\geq0$, $\nu\geq0$, $\lambda_1>0$ and $\lambda_2>0$ are fixed parameters, μ controls the smoothness of zero level set, ν increases the propagation speed, and $\lambda_1$and λ_2 control the image data driven force inside and outside the contour, respectively. $c_1$and c_2 are two constants which are the average intensities inside and outside the contour, respectively.

$$c_1(x)=\frac{\int_\Omega I(x)H(\phi)dx}{\int_\Omega(1-H(\phi))dx} \tag{4}$$

$$c_2(x) = \frac{\int_{\Omega} I(x)(1 - H(\phi))dx}{\int_{\Omega} (1 - H(\phi))dx} \quad (5)$$

$H(\phi)$ is the Heaviside function given by:

$$H_{\varepsilon}(\phi) = \frac{1}{2}\left(1 + \frac{2}{\pi}\arctan\left(\frac{\phi}{\varepsilon}\right)\right) \quad (6)$$

The new level set model, obtained by the association the two cited models, is formulated as follow:

$$\frac{\partial \phi}{\partial t} = spf(I(x))\left(div\left(\frac{\nabla \phi}{|\nabla \phi|}\right) + \alpha\right)|\nabla \phi| + \nabla spf(I(x))\nabla \phi \quad (7)$$

The SPF function modulates the signs of the pressure forces inside and outside the region of interest. The SPF function is given as follow:

$$spf(x) = \frac{I(x) - \frac{c_1 + c_{\acute{e}}}{2}}{\max\left(\left|I(x) - \frac{c_1 + c_{\acute{e}}}{2}\right|\right)} \quad (8)$$

After many approximations which can be found in [11], the implemented level set formulation of the model is written as follow:

$$\frac{\partial \phi}{\partial t} = spf(I(x)).\ \alpha\ .|\nabla \phi|, \qquad x \in \Omega \quad (9)$$

The principal steps of the algorithm are given in the follow:

Step 1: Initialize the level set function ϕ as

$$\phi(x, t = 0) = \begin{cases} -\rho & x \in \Omega_0 - \partial\Omega_0 \\ 0 & x \in \Omega_0 \\ +\rho & x \in \Omega - \Omega_0 \end{cases}$$

where $\rho > 0$ is a constant, Ω_0 is a subset in the image domain Ω and $\partial\Omega_0$ is the boundary of Ω_0.

Step 2: Compute $c_1(\phi)$ and $c_2(\phi)$ using (4) and (5), respectively,

Step 3: Evolve the level set function according to (7),

Step 4: Let $\phi = 1$ if $\phi > 1$ otherwise, $\phi = -1$.

Step 5: Regularize the level set function with a Gaussian filter, *i.e.* $\phi = \phi * G_{\sigma}$

Step 6: Check whether the evolution of the level set function has converged. If not, return to step 2.

V. TESTING

To test the reliability and robustness of this system, several videos were recorded in different conditions. Each video includes a set of vehicles taken at different lighting and pose scenarios. The length of each video is about three minutes. The background, contrast, lighting, and shadow conditions of these scenarios were quite different. The frames were extracted afterward and stored in a hard disk to be tested later. Different vehicle size classes with different colors were involved in this experiment. Single and multiple vehicle scenarios in a scene have also been considered.

A. Single Vehicle Detection

Figure 2 highlights the performance of preprocessing and level set based segmentation stage of the image shown in Figure 3a. It shows that after application of background subtraction and connected component labeling, the other objects detected in the segmentation were excluded, this corresponds to house, trees of the original image, and only the vehicle image is considered as demonstrated in the last part (Fig. 2f). This demonstrates that the algorithm successfully identified the vehicle in the original image. On the other, the output of the level segmentations at various iterations levels has also been pointed out.

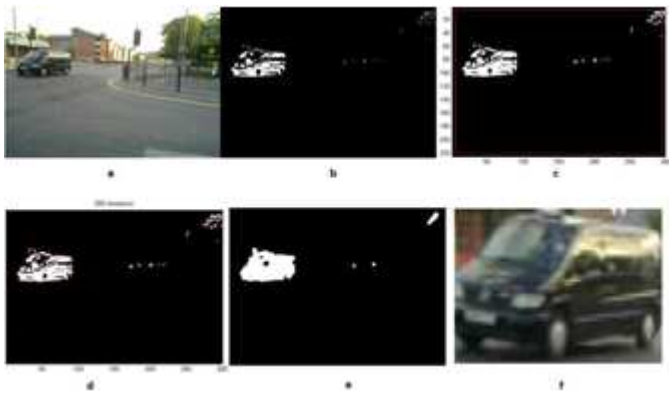

Figure 2. Detection process of a single Vehicle

B. Multiple Vehicle Detection

In this scenario, more than one vehicle is included in the frames. The accuracy of the system is evaluated through the detection performance on all vehicles presented in the frames. Figure 3 shows the detection procedures implemented in this experiment. As it can be noticed all vehicles in the frame have been detected. Although the system is meant to detect the vehicle on the right-hand road in the frame only, it also detected the ambulance car in the top left side of the frame as illustrated in Fig 3f, which testifies of the robustness of the developed system. Similarly to Figure 2 intermediate results of the level set segmentation have also been highlighted.

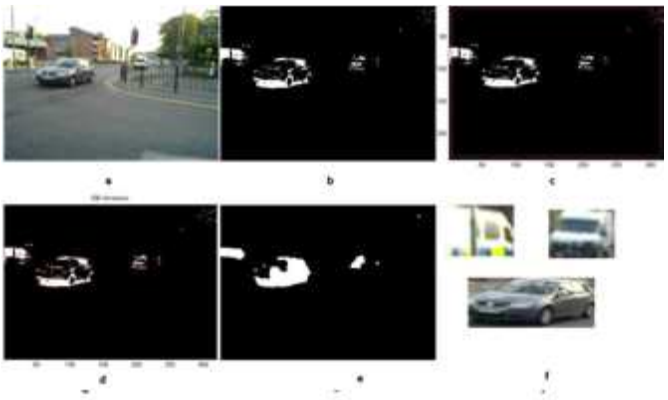

Figure 3. Detection process for detection of multiple vehicles

Nevertheless, there are also scenarios in which there is a missing detection for various reasons, linked mainly to preprocessing stage and quality of images. For instance, in Figure 4 the black vehicle was not detected. This is due to the threshold applied to the system. This case, however, will not affect the performance of the system because in the following frame, this missed vehicle will be detected as shown in Figure 5.

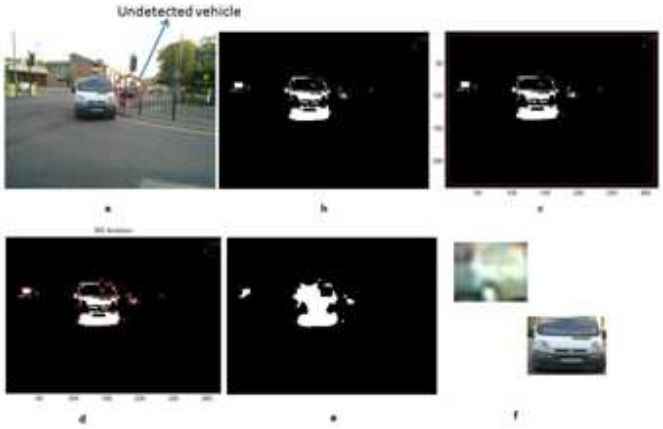

Figure 4. Example of missed/incomplete detection

The performance of this system is being measured using true detection rate. This calculates the number of true detection over the overall number of object.

*True detection rate ( ) = (correct detection) / (the total number of objects) * 100%*

For this purpose, three videos with various number of vehicles have been used to test the performance of the system. The summary of this evaluation is reported in Table 1. It is clear that some of the vehicles in video were not detected by the detection system; thus, they appeared as missed.

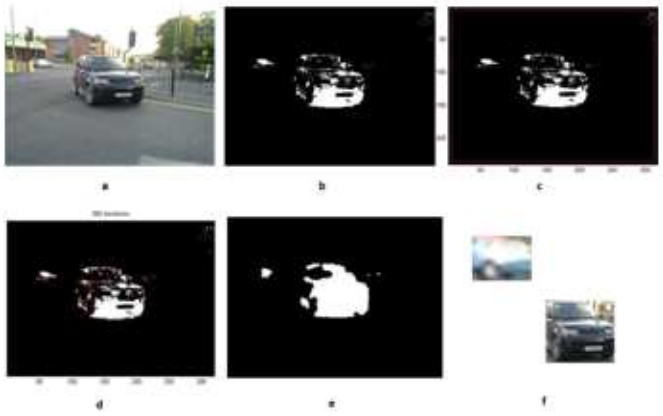

Figure 5: Detection of the missed object (car) of previous frame

Strictly speaking, the main reason for this phenomenon relies on the restriction of a single-camera system. Indeed, in such scenarios it is often employed another camera on the other side of the road that would allow us to detect these vehicles. Nevertheless, despite, the system configuration limitation, the achieved accuracy in terms of detection rate (efficiency) is very much acceptable. Table 1 also indicates the wrong detection corresponding to an object that is wrongly assimilated to a vehicle. For instance, for video 1, this corresponds to 41-32-4 = 5/41 (12.2%)

Table 1. Detection Rate Performance

Frames	Nb. Vehicles	Correct Detections	Miss	Efficiency
Video 1 20 frames	41	32	4	78%
Video 2 15 frames	36	31	5	86.1%

Next, in terms of similarity matching performance, where the purpose is to recognize vehicles that are similar to operator's query image. In this case, the performance of the similarity matching of the various colour features will be evaluated. The experiments will be done on many frames in order to retrieve the similarity between those frames and the operator's query image. Given two frames A and B, the similarity indices F_1, F_2 and F_3 corresponding to color space, dominant color and color layout descriptors, respectively, are given as

$$F_1(A,B) = \sum_H \sum_S \sum_V \frac{\min(A(H,S,V), B(H,S,V))}{\min(|A|, |B|)}$$

In case of dominant color descriptor, given two frames given in term of (c_i, p_i) evaluations; namely, A={ (c_i^1, p_i^1) i=1,N_1}, B={ (c_i^2, p_i^2), i=1,N_2},

$$F_2(A,B)=\sqrt{\sum_{i=1,N_1}(p_i^1)^2+\sum_{i=1,N_2}(p_i^2)^2-\sum_{i=1,N_1}\sum_{j=1,N_2}2a_{1_i,2_j}.p_i^1p_j^2}$$

where $a_{1_i,2_j}$ stand for similarity coefficients between the dominant colors c_i^1 and c_j^2:

$$a_{1_i,2_j}=\begin{cases}1-\left\|c_i^1-c_j^2\right\|/d_{\max} & \text{if } \left\|c_i^1-c_j^2\right\|\le T_d\\ 0, & \text{otherwise}\end{cases}$$

T_d is the maximum distance between two colors to be considered similar, $d_{max}= \alpha T_d$. It is commonly chosen Td in the range [10, 15] interval, while α in [1 1.5].

Finally, when using the color layout Feature, the similarity matching for two CLDs , {*DY, DCr, DCb*} and {*DY, DCr, DCb*}, is computed as

$$F_3(A,B)=\sqrt{\sum_i w_{yi}(DY_i-DY_i')^2}+\sqrt{\sum_i w_{bi}(DCb_i-DCb_i')^2}+\sqrt{\sum_i w_{ri}(DCr_i-DCr_i')^2}$$

Where, *i* denotes the Zig-Zag scanning order of the coefficients.

To evaluate the retrieval performance of these descriptors, the precision and recall indices are compared with respect to the above three similarities. The precision calculated the proportion of good match among those retrieved by the system, while the recall looks at whether there is no missing relevant car among those identified by the system. The result is summarized in Table 2.

Table 2. Precision and Recall evaluations using different color features

Color feature	Nb of trials	Precision	Recall
F1	32	67.74 %	74.2%
F2	32	56.25%	74.6%
F3	32	87.09%	92.4%

From the preceding, one notices that

- The developed approach provides quite good results in terms of detection of the objects of interest, which consists of cars, even in case of strong occultation with multiple cars.
- The efficiency of the retrieval process when comparing the detected cars with the user request car indicates a relatively good performance of he colour layout feature as compared to color space and dominant color features with acceptable performance rate in terms of precision and recall.
- It should be noted that the outcome of the system is strongly related to the background subtraction task. Indeed, it has been observed that when the environment in terms of illumination change is more stable, then the performance of the system increases. That is why the choice of video clips at the beginning is strongly related to this issue, and also, yields less computational cost to eliminate non-target objects.
- The results are also affected by the size of the image, especially when it comes to the similarity matching process using the colour histogram descriptor. This happens due to the structure of the intersection histogram equation. So, when the difference in size between the two vehicles is large, wrong outcomes of this method may occur.
- It should also be noticed that the restricting the analysis to only colour descriptors is another limitation as all texture and other high level features are fully ignored.

REFERENCES

[1] A. Black and M. Isard, *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion*, Springer, 1998

[2] S. Osher, J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. Journal of Computational Physics, 1988,79(1): 12-49

[3] Malladi R, Sethian J A, Vemuri B C., Shape Modeling with front propagation: A level set approach. IEEE Transactions on pattern analysis and machine intelligence, 1995, 17(2):158-175.

[4] M. Kass, A. Witkin, D. Terzopoulos. Snakes: Active Contour Models. International Journal of Computer Vision, 1987, 1(4): 321-331

[5] L. D. Cohen, On Active Contour Models and Balloons. CVGIP:IU,1991,53:211-218

[6] X. Xie, Majid Mirmehdi, RAGS:Region-Aided Geometric Snake,IEEE Transactions on Image processing, 2004,13(5): 640-652

[7] V. Caselles, R. Kimmel, G. Sapiro, "Geodesic active contours, " in Processing of IEEE International Conference on Computer Vision'95, Boston, MA (1995) 694-699.

[8] Paragios N, Deriche R., Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects. IEEE Transactions on pattern analysis and machine intelligence, 2000, 22(3):266-280

[9] T. F. Chan, L. A. Vese, "Active contours without edge," in IEEE Transaction on image processing, Vol. 10, NO. 2, February 2001.

[10] D. Mumford, J. Shah, Optimal Approximation by piecewise smooth functions and associated variational problems. Communications on Pure and Applied Mathematics, 1989, 42:577-685

[11] K. Zhang, L. Zhang, H. Song, W. Zhou, "Active contours with selective local or global segmentation: a new variational approach and level set method," in Image and Vision Computing 28(4), 2010, 668-676.

[12] S.-F. Chang, T. Sikora, and A. Puri, "Overview of the mpeg-7 standard," IEEE trans. on Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 688–695, 2001.

[13] S.-C. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," in Proc. EI-VCIP, 2004, pp. 881–892

[14] J. B. McFarlane and Schofield,C. P. Segmentation and tracking of piglets in images. Machine Vision and Applications. 8(3), 1995, 187-193.

[15] M.B. Dillencourt, H. Samet, M. Tamminen, A general approach to connected-component labeling for arbitrary image representations, J. ACM 39 (2) (1992) 253–280

[16] Motameni H., Norouzi M. Jahandar M. and A. Hatami. Labeling Method in Steganography. World Academy of Science, Engineering and Technology. 2007, 30 (66),pp 354-349

[17] M. Dillencourt, H. Samet, M. Tamminen. A general approach to connected component labeling for arbitrary image representations. Journal of the ACM, 1992, 39(2):253-280.

[18] D. P. Mukherjee, N. Ray, S. T. Acton, Level Set Analysis for Leukocyte Detection and Tracking, IEEE Transactions on Image processing, 2004, 13(4): 562-572